TITLE

Cyberbullying Detection Using Machine Learning

Submitted by:
Chinaza Okeke

Challenge:
CuraJOY Cyberbullying Project – 2025

Date:
June 29, 2025

1. Introduction

Cyberbullying is a modern crisis affecting individuals globally, especially youth on social media. The goal of this project is to develop a machine learning pipeline capable of detecting bullying language in text, offering early detection and prevention. This solution is built as part of the CuraJOY Cyberbullying Challenge, aligning with goals to promote digital safety and empathy online.

2. Dataset & Preprocessing

The dataset consisted of manually labeled social media text samples, annotated as either **"Yes" (bullying)** or **"No" (not bullying)**.

- **Initial size:** ~70 records
- **Expanded to:** 142 (balanced 71/71)

Preprocessing Steps:

- Lowercasing
- Removing special characters
- Stopword removal (using nltk)
- Lemmatization for word consistency

To improve balance, **upsampling** was applied to the minority class.

3. Model Architecture

The final model used was a **Random Forest Classifier**, trained on text transformed via **TF-IDF vectorization**.

Pipeline:

- TF-IDF (max_features=1000)
- RandomForestClassifier with class_weight='balanced'
- Applied a **threshold of 0.3** for better recall in minority class

Why Random Forest?

- Performs well on small datasets
- Handles class imbalance when tuned
- Offers probabilistic output

Other models like Logistic Regression were tried but underperformed on recall.

4. Evaluation Results

Final Determination:

Metric	***Score***
Accuracy	65.5%
Precision	58.8%
Recall	76.9%

F1Score	66.7%

- The model tends to **detect bullying well (high recall)**.
- Still misclassifies subtle phrases like “You are stupid” occasionally.

5. Streamlit Web App

The project was deployed with **Streamlit**, allowing users to enter any sentence and get a prediction.

Features:

- Input box for text
- Output: **Bullying** or **Not Bullying**
- Shows **confidence score**
- Added rule-based check for known bullying phrases (e.g., "idiot", "dumb")

6. Challenges

- **Annotation Time:** Manual labeling was time-consuming.
- **Edge Cases:** Phrases like “You are stupid” were still misclassified.
- **Deployment Issues:** Git, file conflicts, and rebase errors took time.

. Conclusion

This project successfully built a basic bullying detection pipeline using accessible ML techniques. While not perfect, it highlights:

- The importance of **balanced data**
- The impact of **threshold tuning**
- The need for both **ML + rule-based systems**

Future Work:

- Increase dataset size
- Use transformer-based models (e.g., BERT)
- Add multilingual support
- Deploy API for real-time usage

🎯 Final Deliverables:

- Cleaned dataset (CSV)
- Trained model + vectorizer (Pickle files)
- Streamlit App
- GitHub Repo: github.com/chinaza-okeke/CuraJOY-cyberbullying-Project

- This report (PDF format)